Scientometric and Webometric methods

Ingwersen, Peter

Published in:
Tushu Qingbao Zhishi

Publication date:
2010

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Ingwersen, P. (2010). Scientometric and Webometric methods. *Tushu Qingbao Zhishi*, *1*(133), 4-11.

Scientometric and Webometric Methods

Peter Ingwersen
(Royal School of Library and Information Science Birketinget 6, DK 2300 Copenhagen S. Denmark)

[Abstract] The paper presents two fundamental models of scientific communication and characterizes and exemplifies the concept of Scientometrics' and its sub-research areas: publication analysis, including so-called publication point evaluation; citation analysis; and crown indicators for research evaluation. The recent research area Webometrics' is briefly discussed.

1 Scientific Communication Models

Commonly the scientific production in the form of journal articles, conference proceedings papers and research monographs are evaluated by means of peer reviewing prior to publication. Site visits in laboratories by peers may further assure the quality of the research and its communication to the academic community. Another way of dealing with the recognition and, to a certain extent, quality of research outcome is to analyze the citations received by the publications from peers over a given time period. The latter analysis methodologies assumes that only original and peer reviewed publications form part of the analysis, i.e. that both the cited as well as the citing publications are peer reviewed and follow common academic conventions. In the digital age this assumption (or condition) may be difficult to maintain owing to the variation of digital communication platforms presently available. Two models of scientific communication are competing.

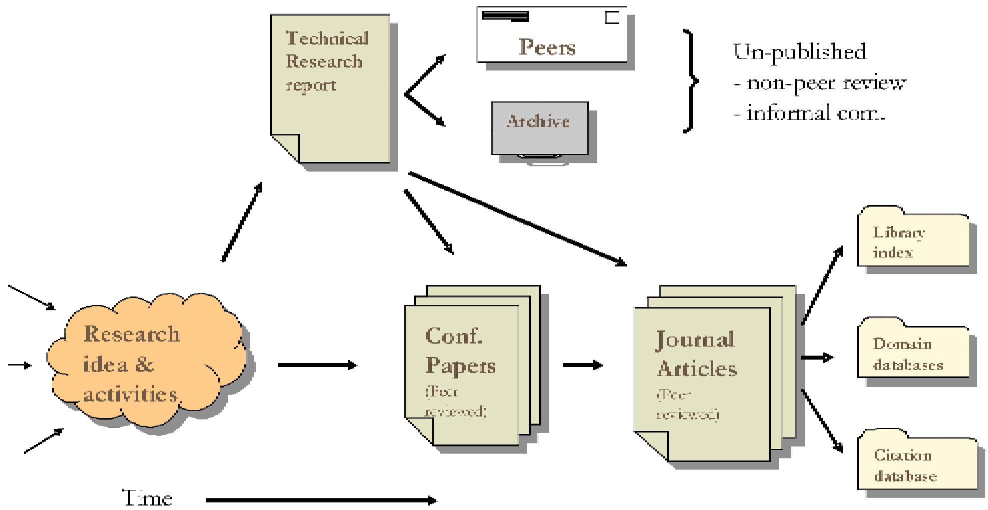

Diagram 1 Classic model of scientific communication

The classic model—Diagram 1—adheres to the pre-web and pre-open access age of scientific communication, i.e. the time prior to 1995. The communication of pre-prints via surface mail was cumbersome. The technical reports were virtually non-accessible, since they were stored as single copies in closed institutional archives. The travel of a publication from research idea to paper or/and journal article was straight forward. Most importantly, access to the full text of the publications was restricted to access to the proceedings and journals proper in their printed versions. Only surrogate records (often with abstracts though) were accessible in libraries, bibliographic and citation databases.

This picture has changed radically in the digital age Diagram 2. The technical reports and working papers are often accessible in the public domain via institutional repositories. Although not peer reviewed such document types are increasing mixing with and difficult to distinguish from peer reviewed material. Secondly, the peer reviewed monographs, conference papers and journal articles are not only accessible in a restricted manner (pay-per-view) from the publisher, conference proceeding and journal websites but also accessible in a likewise restricted manner from full text domain or citation databases, as well as Amazon. com. In addition, some papers and journal articles are freely accessible via Google Scholar or via personal homepages of the scientists or because articles are published in open access journals.

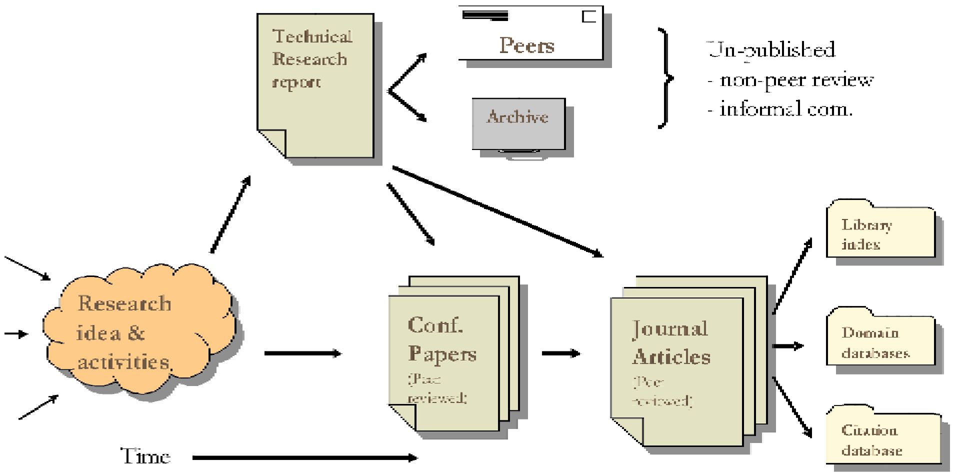

Diagram 2 Present model of scientific communication

Furthermore, the communication landscape has become a mixture of authoritative sources (the peer reviewed ones, Diagram 2, including the open access journals) and less authoritative but opinion-based sources, like student output, work-in-progress materials, academic blogs, wikis, etc. When research evaluation analysis of this landscape is performed great care must be taken to filter out the noise so that only the peer reviewed publications form part of the analysis. Naturally, sources like blogs or working-in-progress papers are indeed legitimate objects for study in themselves as part of the communication process but again: the crucial step is to filter out such objects from the mass of sources available in digitalized form.

Having carried out this filtering process several scientometric analysis possibilities are now available. Scientometrics′ implies quantitative studies of academic output (Bj öneborn & Ingwersen, 2001; 2004). Such studies are commonly divided into publication and citation analyses and may lead to a variety of research evaluation and mapping methodologies. The remaining part of the paper first exemplifies publication analysis. This is followed by characteristics of citation analysis, including the use of Google Scholar and the Hirsh Index, and the application of Crown Indicators in research evaluation. Finally, we discuss briefly selected Webometric indicators.

2 Publication Analysis

Basically, publication analysis signifies to count publications as well as calculating a variety of distributions, e. g. articles over journals or number of authors over articles. Commonly one may count (peer reviewed) publications in

- Academic fields or disciplines
- Authors of communications, like no. of authors per paper
- Countries, regions, universities, departments, research groups
- Communication vehicles, like journals, conference proceedings
- Time periods, in the form of time series
- Mixtures of the above, like no. of journals in a field used by a country

Diagram 3 displays a typical time series of growth, however with an non-typical distribution.

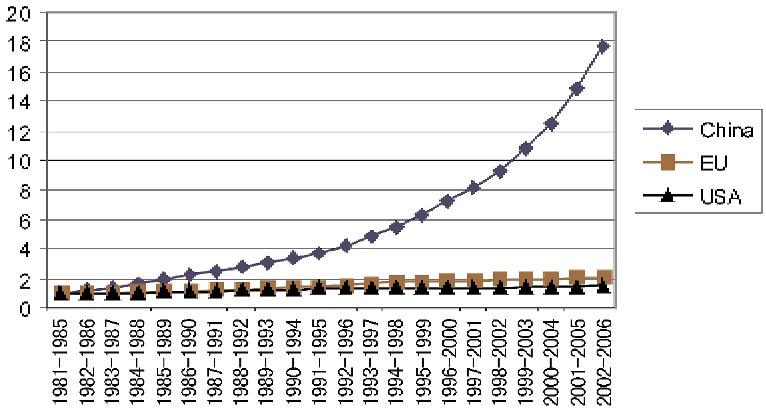

Diagram 3 Publication growth in Thomson-Reuter's Citation databases for China, EU and USA all academic fields, 1981-2006 (National Science Indicators, 2007)

The Index value is set to 1 for the first five-year period in this running time series and we observe that on the surface China's growth seems to follow a power law. However, the diagram demonstrates alone the growth of China's publications indexed by Thomson Reuters in their three citation databases, Social Science Citation Index, Science Citation Index and Arts & Humanities citation Index. That is, the curve for China signifies the growth of China's *international contributions*, as indexed by the citation databases, not the real growth of all Chinese publications including those published nationally.

2.1 Publication Points

In the publication analysis above mainly journal articles are included, i. e. , that the science and biomedical domains are dominating the data. The picture might be different if monographs and conference papers were included too. However, the traditional publication analysis assumes that each publication type counts as equal. Another way of dealing with publication analysis is to apply different weights for different publication types. This is the method applied in the so-called Norwegian publication point system for research evaluation (Sivertsen, XXXX). Journal articles from the 20 % most qualitative international journals in each domain (as judged by peers) receive 3 publication points; other peer reviewed journal articles receive 1 point. Peer reviewed conference papers gets .7 points and monographs from the best' international publishers (assessed by peers) obtain 8 points. Research monographs from other publishers receive 5 points. Points are fractionally distributed between universities and research institutions at author level for each publication. This system also covers the humanities and social sciences and is applied as key for the allocation of public research funding in Norway. Owing to the different publication traditions in the different domains (e.g., many authors per publication in several natural science fields, and very few coauthors in the humanities) the system seems quite fair. The advantage is that the humanities are included, which are difficult fields if citation analysis should be applied instead. On the other hand, the publication points only refer to the success of publication in particular journals (or conferences/publishers) not directly to the quality of the publication or the research itself.

3 Citation Analysis

Citation analyses may be carried out in a *synchronic* way, i.e., observing citations back in time from a given point in time, or in a *diachronic* way, that is, forward in time from a given point in time (Ingwersen et al., 2001). The synchronic analysis type is like done for the ISI Journal Impact factor (JIF). The ISI-JIF is calculated by observing the number of citations given in a specific year (Y) back in time to publications published in a journal the previous two years (Y-1; Y-2). The ISI-JIF hence demonstrates the average impact of the articles published within those two years, with a maximum citation window of 2 years. If applied as a score for the individual articles the ISI-JIF simply signifies the *average publication success* of the article, by being published in a particular journal. It does not infer anything of the *real* impact (or recognition or quality) of the article (Seglen, 1997). One might just as well apply the Norwegian publication point system outlines above which, as argued, also may cover monographs and humanities fields.

The diachronic citation analysis signifies the citation impact of older research as recognized by more recent or current research by means of citations. As for publication analyses one may carry out citation analyses on academic fields, geographical areas, institutions publication types, etc. as outlined above. Time series are frequently used to demonstrate the trends over a longer period of time, see e.g. Diagram 4, displaying a time series for absolute citation impact for USA, EU and China, as observed in the Thomson-Reuters citation indexes. Again, China's curve signifies the *international* development of China's recognition world wide 1981-2006.

Aside from simplistic citation counting ($\sum c$) and absolute impact calculations ($\sum c/\sum p$), where c and p signify citations and publications from a given research entity and time period, Scientometrics also applies indicators *normalized* against some benchmark. Typically, for a specific research field in a unit or a specific journal, one may normalize against the area in a global sense. The formula would hence look like this: $(\sum c/\sum p)/(\sum C/\sum P)$, where C means the volume of citations for the area worldwide and P the corresponding number of publications worldwide.

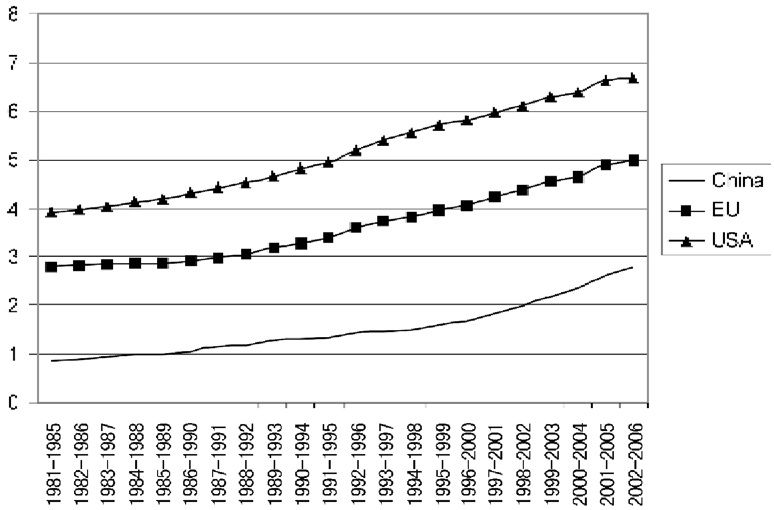

Diagram 4 Absolute citation impact 1981-2006 in Thomson-Reuters' citation databases (2007)

The normalized citation impact is important, because one may observe how a given field in a country perform against the same field in a region or globally see e. g. Diagram 5 for each field (*China c/p*) compared to the *Global C/P* for the same field. Likewise, the diachronic JIF of a journal can be compared to the worldwide research area impact for the same diachronic period, to which the journal belongs. One may hence observe if researchers are publishing in low or high impact journals of a given research area.

3.1 The Hirsch Index and Google Scholar

The *H*-index was invented fairly recently (Hirsch, 2005) and combines publication volume and citation distribution into a composite index for a given unit over a longer span of time, for instance, a scholar, a research group or an institution. *H* is the number of articles receiving a number of citations larger or equal to *H*. So, if *H* is 13 for a scholar signifies that he or she has published 13 items that each has obtained *at least* 13 citations. Obviously, the common pattern is that one article has received 13 citations and 12 items have received many more citations. The remaining publications by the person (or group or institution) have received 13 or less citations. Many derivations of the *H*-index are developed, e. g. the *G*-index (Egghe, 2006). The indexes are dependent on research field (citation density) and the time period applied to the calculation, e. g. the life span of a person or a limited number of years.

The open access citation database Google Scholar (www. scholar. google. com) provides access to academic and scholarly-like publications harvested from the Web and sorted by citation volume. The most easy search strategy is to enter a personal name and retrieve that person's publications in descending citation order. As such Google Scholar provides a crude *H*-index, but care must be taken to clean up the output prior to establishing the final citation observations owing to many versions and duplicates of the same papers both on the citing and cited side. Publish or Perish constitutes a help to analyze Google Scholar output in a highly structured way that helps getting control of the output (www. Harzing.

com/pop. htm). Google Scholar is useful in conference paper rich research fields, like engineering and computer science, since the proceedings often are put on to the Web.

3.2 The Crown Indicators

The most robust citation indicators, the Journal Crown Indicators (JCI) and Field Crown Indicators (FCI), are necessary when comparing countries or institutions, like universities or research groups van Raan, 1999). The JCI is calculated as the ratio between the mean number of actual citations received for all the journal articles published by a given unit and the mean diachronic citation impact of the same journals used by the unit over the same time period. The latter impact signifies the expected impact of the unit, having used a specific set of journals.

The FCI is more complicated Diagram 5. The idea behind FCI is the same as for JCI, namely to assess a given unit in terms of its research profile. Where JCI is based on the unit's journal profile′ FCI is based on a unit's research field profile′. What is done when calculating FCI is to create a shadow unit′ with the same profile, but at *world level*. Diagram 5 demonstrates the procedure for China.

China, Research Profile 2001-05		**China c/p**	**Cits**	**Publ.**	***Profile%***	**Global C/P**	**Indicator**	**Weighted Cits.**
Agricultural & Plant Sc.	2001-2005	2.19	21137	9653	4.18	2.89	0.76	27897.17
Biology & Biochemistry	2001-2005	3.66	30563	8352	3.62	7.56	0.48	63141.12
Chemistry	2001-2005	2.89	155384	53851	**23.34**	4.28	0.68	230482.28
Clinical Medicine	2001-2005	4.28	75917	17736	7.69	5.4	0.79	95774.4
Computer Science	2001-2005	1.17	3821	3256	1.41	1.51	0.77	4916.56
Ecology/Environment	2001-2005	2.26	10907	4835	2.10	3.59	0.63	17357.65
Engineering	2001-2005	1.47	35873	24463	**10.60**	1.78	0.83	43544.14
Geosciences	2001-2005	2.54	19422	7655	3.32	3.44	0.74	26333.2
Immunology	2001-2005	4.33	3976	918	0.40	10.62	0.41	9749.16
Materials Science	2001-2005	2.07	40779	19684	8.53	2.54	0.81	49997.36
Mathematics	2001-2005	1.1	8542	7732	3.35	1.32	0.83	10206.24
Microbiology	2001-2005	4.19	10027	2394	1.04	6.9	0.61	16518.6
Molecular Biology & Genetics	2001-2005	7.26	18395	2533	1.10	12.63	0.57	31991.79
Multdisciplinary	2001-2005	2.24	14650	6531	2.83	4.48	0.50	29258.88
Neurosciences & Behavior	2001-2005	4.37	10384	2374	1.03	7.88	0.55	18707.12
Pharmacology	2001-2005	2.47	8963	3633	1.57	5.01	0.49	18201.33
Physics & Space Sc.	2001-2005	2.56	138329	53109	**23.02**	4.12	0.62	218809.08
Social Sciences general	2001-2005	1.65	3245	1971	0.85	1.99	0.83	3922.29
Ratio/Sum		**2.65**	610314	230680	100	4.69	0.56	916808
		(Weighted) Field Crown Indicator: 610314/916808.37=						***0.67***

Diagram 5 Research profile of China 2001-2005 and the corresponding normalized Field Crown Indicators for each field and the weighted mean FCI (National Science Indicators, 2006)

China's research profile across all fields except the humanities 2001-05 can be observed in the column '*Profile %*'. One observes that Chemistry, Physics & Space Science as well as Engineering are the predominant fields. The global research profile is quite different from that of China: Clinical Medicine and Biology & Biochemistry as well as Molecular Biology & Genetics are much more dominant fields. If the Chinese research profile was *not* taken into account China's international citation impact would be based on the global one, i. e., on the global dominant fields. If done like that the common normalized citation impact index for China would be 2.65 / 4.69 = 0.56 see bottom row, Diagram 5. This would be unfair to each country since each country's profile is different from the average global one. Instead the FCI calculation is based on the weights provided by each research field of the country. For

China one observes that the FCI for each field (column labeled Indicator) is below index value 1, but in some fields, like Engineering and Mathematics the index values are above .83, whilst in other fields it is quite low, below .50, e.g., Biology & Biochemistry, Immunology, and Pharmacology. However, when adding up all the research fields the citation density of the fields plays a central role. The FCI calculation is done by the following formula: $\Sigma c / \Sigma(C/P^{field}\ x\ p^{field})$ – where C and P^{field} signifies the global citation and publication impact of a given research field (column '*Global C/P*') that is multiplied by the country's publication number in the same field (p^{field}). Essentially, for each field one calculates the number of citations it should have received according to the global impact of that field and the actual number of publications produced by the country. For China, Diagram 5 demonstrates in its right-most column this '*Weighted Cits*' estimation which, when summed up, gives a total of 916,808 citations. This number is the expected number of citations for a global shadow country identical to China, given the volume of China's publications across all its research fields. What is to be done is to divide the actually received citation volume by the expected number of citations to obtain the true FCI (610,314/916,808.37=.67). We observe that in this case China's FCI is quite higher than the simplistic normalized citation impact with no weights applied (.56). This is because some Chinese citation dense fields actually have received a substantial number of citations hereby pushing up the FCI score. This FCI score may now be compared to other countries' FCI scores, since they too are calculated based on *their* research profiles, not the average global one.

4 Webometric indicators

At first, the Web Impact Factor (WIF) was naively believed to be like a citation impact factor, demonstrating at least something about recognition (Ingwersen, 1998). This idea was build on the notion of Webometrics, coined by Almind & Ingwersen in a previous publication on analyses of the young Web (1997). The WIF was calculated by the ratio of received links (inlinks) to a Website divided by the number of pages in the site. However, it became soon clear that most links are navigational and has nothing to do with recognition, not to speak of quality of web pages. Inlinks are *not* similar to citations in that no conventions exist as to their construction as existing in the academic domains for citations.

Björneborn & Ingwersen (2001; 2004) sum up the conceptualizations of Webometrics and how to carry out several indicator calculations. It seems now better simply to apply the number of inlinks to a given Website (minus self-links within the site) as a reasonable measure of recognition or impact of Web-entities (Thelwall et al., 2004). The indicator value seems to correlate with academic production, i.e., the more you produce and add to your Website, the more external inlinks you obtain. The same research review and later studies also imply that it seems justified to divide the external inlink number by the staff volume of a given unit, in order to obtain a reasonable ratio that correlate with other research evaluation scores. There may not be real correlations between a journal IF and the same journal Website WIF, simply because academic entities do not receive many inlinks (compared to citations). It is of course quite interesting in a scientific sense that any PageRank algorithm that sorts retrieved output according to iterative and self-re-enforced scores does that based on the wrong conception that links are like citations. There exist other interesting Impact factors on the Web, like the WUF, We Use Factor, which counts outlinks. Outlinks signify some kind of usage of the outlinked page, i.e., a kind of knowledge import (Thelwall, 2003)

One should also point to associated Web-based indicators, such as number of visits to sites, searches carried out on Websites, volume of downloads done from sites or number of cases of social tagging done to specific entities. These indicators all belong to *social utility* measures of the Web.

5 Conclusions

The paper has outlined the basics of publication analysis, citation impact calculations and the central use and advantages of Crown Indicator over common un-weighted but normalized impact. It pointed to the generation of the *H*-index and derivates and how Google Scholar may be used to create such indexes. Finally, the paper contributes to the understanding of some central Webometric indicators, such as the Web Impact and Use Factors, as well as to other social utility measures applied to the Web.

References

1 Almind TC & Ingwersen P. Informetric analyses on the World Wide Web: methodological approaches to Webometrics. J. Doc., 53(4), 1997: 404-426.
2 Björneborn L & Ingwersen P. Perspectives on webometrics. Scientometrics, 2001, 50(1): 65-82.
3 Björneborn L & Ingwersen P. Towards a basic framework for webometrics. J. Am. Soc. Inform. Sci. & Tech., 2004, 55(14): 1216-1227.
4 Egghe L. Theory and practice of the g-index. Scientometrics, 2006, 69(1): 131-152.
5 Hirsch JE. An index to quantify an individual's scientific research output. PNAS, 2005, 102(46): 16569-16572.
6 Ingwersen P. The Calculation of Web Impact Factors. J. Doc., 54(2), 1998: 236-243.
7 Ingwersen P, Larsen B, Rousseau R & Russell J. The publication-citation matrix and its derived quantities. Chinese Sci. Bull. (in Chinese), 46(8), 2001: 700-704.
8 Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ 1997, 314: 498-502.
9 Sivertsen G. A bibliometric model for performance based budgeting of research institutions. Lecture presented at the 9th International Conference on Science & Technology Indicators, 7-9 September 2006, Leuven, Belgium.
10 Thelwall M. Web use and peer interconnectivity metrics for academic web sites. J. Inf. Sc., 29(1): 1-10.
11 Thelwall M, Vaughan L & Björneborn L. Webometrics. Ann. Rev. Inf. Sc. & Tech., 2004, 39: 81-135.
12 van Raan AFJ. Advanced bibliometric methods for the evaluation of universities. Scientometrics, 1999, 45(3): 417-423.

(收稿日期:2009-12-16)

作者简介:彼得·英格沃森(Peter Ingwersen)博士,1973年毕业于丹麦哥本哈根皇家情报学院,丹麦哥本哈根皇家图书情报学院教授,国际著名情报学家。

彼得·英格沃森博士1991年获得哥本哈根商学院信息管理系博士学位;1993年7月起,任信息检索理论系的系主任;1997年起,他成为芬兰埃博学术大学信息研究系的全职教授;1992—2002年,在北欧研究院赞助的坦佩雷大学情报系(现已更名为NORDFORSK)任客座教授;2006年开始,任丹麦哥本哈根皇家情报学院信息检索与查寻方面的全职教授;世界著名杂志《JASIST》、《Information Processing & Management》、《Scientometrics》等的编委。

彼得·英格沃森博士的主要研究领域包括:联机文献计量学方法、基于任务的交互式IR、IR评估与设计方法。他在数字化管理、信息科学、信息查寻与信息行为、知识组织与信息构建、搜索引擎和用户界面、认知信息检索等方面也有很高的造诣。出版著作多部,其中《The TURN: Integration of Information Seeking and Retrieval in Context》和《Information Retrieval Interaction》两部已受到国际认可;发表学术论文70余篇。